

ユーザの閲覧履歴による Web 検索のパーソナライズ化

目木信太郎[†]

[†] 岡山県立大学情報工学部 〒719-1197 岡山県総社市窪木 111

E-mail: †meki@cse.oka-pu.ac.jp

あらまし インターネット上には多数の WWW サーバが存在するので、多くの利用者は検索エンジンを利用して必要な情報が存在する WWW サーバの URL を取得しているが、検索エンジンは非常に多くの検索結果を返すので、利用者は必要とする情報を見つけられないことも多い。そこで、検索条件を追加して検索結果を絞り込むことが必要となるが、その場合に、利用者が明示的にキーワードを追加せずに検索結果を絞り込む方法として、利用者の過去の閲覧履歴に基づいて検索条件を追加する方法が考えられる。この方法は利用者が事前に何らかの操作をする必要がなく、また個人の情報に基づいて行うので各利用者の関心が多様であっても対応できると考えられる。

キーワード 検索エンジン、リランキング

Shintaro MEKI[†]

[†] Faculty of Computer Science and System Engineering, Okayama Prefectural University

111 Kuboki, Soja, Okayama, 719-1197 Japan

E-mail: †meki@cse.oka-pu.ac.jp

1. はじめに

今やインターネットは高度情報化社会に欠かせないインフラとなった。インターネット上の多くの情報は WWW サーバ上に存在するが、利用者は必要とする情報がどの WWW サーバに存在するのかという位置情報 (URL) を知っていなければ、これらの情報を利用することはできない。

そこで、多くの利用者は検索エンジンを利用して必要な情報が存在する WWW サーバの URL を取得している。このとき利用者は検索エンジンにひとつのキーワードしか入力しないことが多いが、インターネット上には膨大な数の WWW サーバが存在するので、検索エンジンは非常に多くの検索結果を返し、そのため利用者は必要とする情報を見つけられないことも多い。

そこで、検索条件を追加して検索結果を絞り込むことが必要となる。その場合に、利用者が明示的にキーワードを追加せずに検索結果を絞り込む方法として、何らかの情報をもとに検索条件を追加する方法が考えられる。

このとき利用できる情報としては、事前に利用者に記入してもらったアンケート、年齢、性別、職業などの利用者の個人情報、利用者の過去の Web ページの閲覧履歴などがある。

これらの情報を利用して検索の精度を上げることを考えるのであるが、その際に必要となる検索結果を評価する指標として適合率と再現率がある。

適合率 検索結果のうちで、実際に利用者が求めようとしているものの占める割合

再現率 利用者が求めようとしたもののうち、検索結果に含まれるものの占める割合

適合率が高いということは、検索結果には利用者が求めようとしているものが多く含まれているということなので、効率よく情報を見つけることができるということである。また、再現率が高いということは、求めようとしているもので検索結果からもれているものが少ないということなので、求めようとしている情報を見落とすことが少ないということである。

上に挙げた三つの情報のうち、アンケートは事前に利用者により多くの質問に答えてもらう必要があるが、それはあまり現実的ではない。また、個人情報は利用者の関心のあることをある程度絞り込むのには使えるが、同じ年齢、性別、職業であっても、例えば趣味のように、まったく異なることに関心がある場合もあるので、あまり適切ではないと考えられる。

それに対して、過去の閲覧履歴を利用する方法は、利用者が事前に何らかの操作をする必要がなく、また個人の情報に基づいて行うので各利用者の関心が多様であっても対応できると考えられる。

利用者の閲覧履歴は一般に URL のリストとして残るので、URL を解析することで、利用者が閲覧した Web ページに関する情報が得られる。例えば jp ドメインには co.jp, or.jp, ac.jp, ne.jp, go.jp などのサブドメインがあり、それぞれ企業、一般

の組織、教育機関、プロバイダ、政府機関をあらわしている。

したがって、URL を解析した結果 co.jp ドメインにある Web ページをよく閲覧しているのであれば、その利用者は企業が発表するプレスリリースやカタログのような情報を好むということがわかるし、ne.jp ドメインにある Web ページをよく閲覧しているのであれば、個人が公開している日記やエッセイ、趣味などのような情報を好むということがわかる。

したがって、これらの情報を用いれば検索結果を絞り込むことができるが、そのようにして利用者の求めている情報を類推して検索する場合には、たとえそれが過去の閲覧履歴に基づくものであっても、検索結果が利用者の求めているものから外れてしまうこともありうる。すなわち、適合率を上げるために再現率を下げるリスクを取っていることになる。このことは利用者の求めている情報を類推して検索する以上不可避のことであるので、そのリスクに関して利用者に何らかの情報を提供することが必要であると考えられる。

本稿は以下のように構成されている。次の章では、基本的なアイデアについて述べる。第 3 章では、検索結果として得られた URL を使って、本当に URL を解析することによって、Web ページの内容を推測することが可能かどうか、検討を行う。最後の章は、まとめである。

2. 基本的なアイデア

インターネット上で Web ページを特定する場合、普通 URL(Uniform Resource Locator) が用いられる。例えば、

`http://www.rakuten.co.jp/index.html`

などがそうである。この URL は、最初の「http」が Hyper Text Transfer Protocol、つまり HTML 文書を取得するプロトコルを用いることを表しており、「www.rakuten.co.jp」が Web サーバの FQDN(Fully Qualified Domain Name) を表しており、最後の「index.html」が Web ページの、Web サーバのルートディレクトリからの相対パスを表している。

Web ブラウザで Web ページを閲覧する場合、最初の部分は「http://」と決まっているので、この部分からはどのような Web ページを閲覧するのかをうかがい知ることはできない。しかし、Web サーバの FQDN の部分は、上から与えられるのではなく Web サーバのシステム管理者が申請して取得するものなので、そこには Web サーバのシステム管理者の何らかの意図が反映されていると考えられる。

上で挙げた例で考えると、トップドメインの「.jp」は、日本のドメインを表しており、したがって、その下にある Web ページは日本語で書かれていることが多いと考えられる。

また、そのサブドメインの「.co.jp」は、日本で登記された企業しか登録できないので、その下にある Web サーバは、日本企業のものであると考えることができる。

さらに、そのサブドメインの「rakuten.co.jp」は、日本の企業である楽天の Web サーバのシステム管理者が申請して、登録されたものであるはずだから、当然楽天の管理するサブドメインであると考えられる。

最後に、一番左の「www」は、ホスト名であり、システム管

理者がそのホストが担っている機能に応じてつけたと考えられるから、その名前から、そのホストがどんな働きをしているかをうかがい知ることができるはずである。「www」は、おそらく World Wide Web の略であろうから、多分 Web サーバであるだろうということが分かる。

一方、Web サーバの FQDN の右の「/index.html」という部分は、Web サーバのルートディレクトリからの相対パスであるから、この場合は、ルートディレクトリにある「index.html」という名前の Web ページであるということが分かる。

また、この「index.html」という Web ページのファイル名は、Web サイトのトップページにつけられることが多いから、この「http://www.rakuten.co.jp/index.html」という URL は、日本の企業である、楽天の Web サイトのトップページを指し示している、ということが分かる。

これらのことから、Web ページの URL のほとんどの部分には、Web サイトのシステム管理者の意図が反映されていることがわかる。Web サイトのシステム管理者は、後々のシステム管理のしやすさを考慮して、素直な名前を付けるはずであるから、その意図を読み取るのは容易であると考えられる。

以上、「http://www.rakuten.co.jp/index.html」という URL を使って、URL を解析することによって、その Web ページがどのような Web ページなのかが大体分かるということを示した。

3. 検 証

前の章で、「http://www.rakuten.co.jp/index.html」という URL を使って、それからどのような Web ページなのかを推測することができることを示した。

この章では、もっといろいろな URL を使って、その URL が指し示す Web ページがどのような Web ページなのかを推測することができるかどうかを検証することにする。

URL を得るために、2012 年 1 月 12 日の Google [1] の人気急上昇キーワードの第一位である「優香」というキーワードで検索を行い、検索結果として得られた URL について、検証を行うことにした。

検索結果の第 1 位の URL は「ja.wikipedia.org」という Web サーバの FQDN を含んでいるので、日本語で書かれた Wikipedia の Web ページである、ということが分かる。

検索結果の第 2 位の URL は「yuka-club.jp/」である。この URL からは、残念ながら日本にある Web サイトのトップページである、ということしか分からない。

検索結果の第 3 位の URL は「pc.yuka-club.jp/blog/」である。Web ページを指し示す Web サーバのルートディレクトリからの相対パスに、「/blog/」というディレクトリを含んでいるので、この URL が指し示すのは、ブログの Web ページである、ということが分かる。

検索結果の第 4 位の URL は「netaatoz.jp/archives/6585197.html」である。この URL も Web サーバのルートディレクトリからの相対パスに、「/archives/」というディレクトリを含んでいるので、おそらく

(1) 優香 - Wikipedia

優香（ゆうか、本名:岡部 広子（おかべ ひろこ）、1980 年 6 月 27 日 - ）は、日本のマルチタレントで、元グラビアアイドル。所属芸能事務所は、... 芸名を週刊プレイボーイ（1997 年 11 月 25 日発売号）やインターネットで一般公募し、「優香」となった。由来は、当時の ...

ja.wikipedia.org/wiki/優香

(2) 優香くらぶ

ホリプロによる公式サイト。新着情報、活動記録、ファンクラブの案内。

yuka-club.jp/

(3) Blog - 優香くらぶ

メッセージを読んできると、『子供の名前を優香と名付けました』という方が多くいらっしゃることにびっくりです。『バカ殿様の優香姫が大好き』という、小さい子達もたくさんいて。『笑顔が大好きです!』『元気がでます!』『癒されます♪』『優香ちゃんになりたい!』...

pc.yuka-club.jp/blog/

(4) 【画像あり】最新の優香のおっぱいがヤバい wwwwww — ねた AtoZ

47 分前 ... タレントの優香（31）が 12 日発売の女性誌「FRaU」（講談社）の表紙で大胆ショットを披露している。8 カ月間にわたるトレーニングや食生活の見直しで、体重 7 キロ、ウエストは 10 センチ減に。ショートパンツ姿で、ウエストのくびれを強調した ...

netaatoz.jp/archives/6585197.html

(5) 【画像】優香の横乳エロすぎ wwwwww - スポン!

1 時間前 ... 【画像】優香の横乳エロすぎ wwwwww. ... タレントの優香（31）が 12 日発売の女性誌「FRaU」（講談社）の表紙で大胆ショットを披露している。8 カ月間にわたるトレーニングや食生活の見直しで、体重 7 キロ、ウエストは 10 センチ減に。...

spon.me/archives/51731349.html

(6) 【やっぱり優香かわいいよな速報】優香（31）大胆ショット披露

45 分前 ... アルファルファモザイクの【やっぱり優香かわいいよな速報】優香（31）大胆ショット披露についての詳細情報。

alfalfalfa.com/archives/5077680.html

(7) 優香 大胆ショット披露「今のカラダが一番好き」 — スポニチ Sponichi ...

44 分前 ... タレントの優香（31）が 12 日発売の女性誌「FRaU」（講談社）の表紙で大胆ショットを披露している。8 カ月間にわたるトレーニングや食生活の見直しで、体重 7 キロ、ウエストは 10 センチ減に。ショートパンツ姿で、ウエストのくびれを強調した背中 ...

www.sponichi.co.jp/entertainment/.../K20120112002411510.html

(8) 優香 - YouTube

2011 年 4 月 17 日 ... Sign In or Sign Up now to post a comment! 健康なお色気で、改めて優香さんのもつ自然体から醸し出る？フェロモンを思いっきり吸い込んでしまいました、。水着姿がとてもお似合い すねえ、... それにしても、オオキイですね、。...

www.youtube.com/watch?v=gwEZL9xeU_o

(9) 優香、艶めかしい S 字ラインを大胆披露 ~ 女性誌企画で 7kg 減の - Yahoo!

33 分前 ... Yahoo!ニュース（オリコン） - 女優でタレントの優香（31）が 12 日発売の女性誌『FRaU（フラウ）』（2 月号 / 講談社）

headlines.yahoo.co.jp/hl?a=20120112-00000312-oric-ent

(10) 優香、艶めかしい S 字ラインを大胆披露 ~ 女性誌企画で 7kg 減の ...

2 時間前 ... 女優でタレントの優香（31）が 12 日発売の女性誌『FRaU（フラウ）』（2 月号 / 講談社）のダイエット企画で 7kg の減量に成功し、より磨きのかかった艶やかなボディを大胆に披露している。30 代に突入した優香だが、「今のカラダが一番好き」と自負 ...

beauty.oricon.co.jp/news/2005638/full/

はニュース的な記事が書かれた Web ページであろう、ということが推測できる。「6585197.html」という Web ページのファイル名は、多分 Web ページのシリアル番号であろう。

検索結果の第 5 位の URL は「spon.me/archives/51731349.html」である。この URL は、第 4 位の URL と同じ構造をしており、同じようにニュース記事の Web ページを指し示しているものと考えられる。

検索結果の第 6 位の URL は「alfalfalfa.com/archives/5077680.html」である。この URL もまた同じ構造をしており、同じようにニュース記事の Web ページを指し示していると考えられる。

検索結果の第 7 位の URL は「www.sponichi.co.jp/entertainment/.../K20120112002411510.html」である。この URL は、Web サーバの FQDN が「www.sponichi.co.jp」であるので、スポニチというスポーツ新聞を発行している新聞社の管理している Web サーバである、ということが分かる。だから、この URL が指し示す Web ページは、当然新聞記事の Web ページであろう。「K20120112002411510.html」の部分は、記事のシリアルナンバーと考えられる。

検索結果の第 8 位の URL は「www.youtube.com/watch?v=gwEzL9xeU_o」である。この URL は、Web サーバの FQDN の部分が「www.youtube.com」であるので、このことから YouTube が提供している、動画の Web ページであるということが分かる。

検索結果の第 9 位の URL は「headlines.yahoo.co.jp/hl?a=20120112-00000312-oric-ent」である。この URL は、Web サーバの FQDN の部分に「.yahoo.co.jp」という部分を含んでいるので、日本のヤフーが提供している Web ページであろう、ということが分かる。また、ホスト名が「headlines」であることから、それはニュースが書かれた Web ページであるだろう、ということも分かる。

検索結果が第 10 位の URL は「beauty.oricon.co.jp/news/2005638/full/」である。この URL もまた同様に、Web サーバの FQDN が「.oricon.co.jp」という部分を含んでいるので、日本のオリコン・エンタテインメントという企業が提供している Web ページであるだろうということが分かるし、Web サーバのルートディレクトリからの相対パスに「/news/」という部分を含んでいるので、それはニュース記事が書かれた Web ページであろう、ということも分かる。

以上、「優香」を検索キーワードにして検索を行った結果得られた URL について検討を行い、その Web ページがどのような Web ページなのかについて推測を行った。検討の結果、一つの URL については、日本にある Web サイトのトップページである、ということしかわからなかったが、それ以外の 9 つの URL については、それが指し示す Web ページの内容を推測することができた。したがって、実際の Web ページを閲覧しなくても、妥当な推測を行うことができたので、本手法は有効であると考えられる。

4. ま と め

2 章で本稿で提案する手法で、なぜ Web ページの内容を推測することができるのかを説明し、3 章で実際の URL を使ってそれを検証した。その結果、URL を解析して、検討を加えるだけでも Web ページの内容を類推することができることを示した。

したがって、ユーザの閲覧履歴の URL を解析すれば、ユーザの嗜好を類推することは十分可能であり、それを利用すればよりユーザの好みに合った検索結果を提示することができるものと考えられる。

文 献

- [1] Google, <http://www.google.co.jp/>